REPORT DOCUMENTATION PAGE

AFRL-SR-AR-TR-05-

0035

1. REPORT DATE *(DD-MM-YYYY)* January 31, 2005	**2. REPORT TYPE** Final Performance Report	**3. DATES COVERED** *(From - To)* May 2001 - October 2004

4. TITLE AND SUBTITLE Defending Against Novel Information Attacks: Prototype Development and Analysis	**5a. CONTRACT NUMBER** **5b. GRANT NUMBER** F49620-01-1-0346 **5c. PROGRAM ELEMENT NUMBER**
6. AUTHOR(S) Thomas Wiggen and Brajendra Panda	**5d. PROJECT NUMBER** **5e. TASK NUMBER** **5f. WORK UNIT NUMBER**
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) University of North Dakota (UND) and University of Arkansas (with a subcontract from UND)	**8. PERFORMING ORGANIZATION REPORT NUMBER** 4041-0701-2004
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Office of Scientific Research 4015 Wilson Blvd., Room 713 Arlington, VA 22203-1954 NM	**10. SPONSOR/MONITOR'S ACRONYM(S)** AFOSR **11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release,
distribution unlimited

Approved for public release,
distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The objective of this project was to develop appropriate techniques for defending information systems from various types of attacks. On intrusion detection front, we developed non-signature based attack detection mechanisms in order to protect information integrity in systems. These developed methods used Petri-Net and Data Mining techniques to identify various attacks. Moreover, we also developed a generalized model for understanding different types of computer viruses in order to develop techniques to protect information systems from them. In addition to attack detection methods, appropriate response techniques were developed that, when applied after discovery of an attack, would aid in bringing the affected system into normal operating conditions.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Thomas Wiggen
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**		6	**19b. TELEPHONE NUMBER** *(include area code)* (701) 777-3477

Final Report

for

DEPSCoR Project

Submitted to

The Air Force Office of Scientific Research

by

The University of North Dakota

Proposal Title: Defending Against Novel Information Attacks: Prototype Development and Analysis

Grant Number: F49620-01-10346

Principal Investigator: Dr. Thomas Wiggen

Major Accomplishments: The following key research activities have been accomplished as a result of this project.

1. Transaction semantic analysis for profiling transactions
2. Identification of malicious transactions in a database system
3. Use of data mining technique for database intrusion detection
4. Mining data relationships for database damage assessment
5. Bit matrix structure for damage assessment in a centralized system
6. Distributed and parallel damage assessment using pre-developed bit matrix structure
7. Damage assessment models for distributed database systems
8. Hybrid log segmentation for databases suffering frequent attacks
9. Fuzzy recovery model for critical information systems
10. Fuzzy dependencies in databases and their applications in damage assessment and recovery
11. Damage assessment and recovery using information flow model
12. Computer virus classification and Network viruses

Executive Summary

Analysis of operating system log or the application log by host-based intrusion detection systems to detect misuse or anomalous activities does not work well for detecting intrusions into a database system. We have developed a method for database intrusion detection by performing

transaction semantic analysis. Typically before a data item is updated in a database some other data items are read or written by the same transaction. After an update, other data items may also be written by that transaction. These data items read or written in the course of updating a data item construct the read set, pre-write set, and the post-write set for that data item. These sets actually reflect the data dependency observed in the transaction.

We have developed the notions of static semantic analyzer and dynamic semantic analyzer in order to generate the read, pre-write and post-write sets. We analyze these sets to determine the extent to which one data item depends on other data items. User applications and database logs can be checked to construct these sets. The static semantic analyzer decides the data dependency by analyzing the transactions statically. The dynamic semantic analyzer observes the extent to which one data item depends on other items by analyzing the database log. These sets are then compared with data items read or written by each transaction. Any deviation from normal update patterns indicates anomalous activities.

Sometimes, users execute a series of transactions to perform a specific task. So, in such a case, analysis of individual transactions may generate false positives. In order to identify malicious user tasks, we use data dependency relationships among transactions and use Petri-Nets to model normal data update patterns at user task level. Dependencies are determined by using the read, pre-write, and post-write sets of data items which are generated by the static semantic analyzer as discussed above. User applications program can be analyzed to construct these sets. We developed notions of write-chain and data update dependency graph (DUDG) to capture normal data update sequence at the user task level. Petri-Nets were used to implement the DUDGs. Since by using the Petri-Nets the partial orders among normal data update operations in a user task can be identified, this method provides a clear and compact model of the data dependency theory and can be executed to identify malicious transactions. In order to choose a reasonable range in the database log to apply the Petri-Net model appropriate techniques have been developed. By using the time pattern of transaction execution, a range of the database log can be calculated. In this range, the data access sequence in a user task is compared with the Petri-Net modeled normal data update sequence. Any deviation from the normal update sequence indicates the existence of a malicious user task. This approach makes the comparison procedure efficient and also reduces false negatives.

We also developed a data mining approach for detecting malicious transactions in database systems. Our approach concentrates on mining data dependencies among data items in a database. Data dependency rules discovered by the data dependency miner are employed as classification rules for identifying anomalies. The experiment on synthetic database transactions illustrated that the proposed method works effectively for detecting malicious transactions in database systems provided certain data dependencies exists. The result further showed that the stronger the data dependency among data items, the better the overall performance.

Having the knowledge of the damage profile after detection of an attack on a database is crucial for the intrusion response team to design an appropriate response strategy. Given the existing approaches to database damage assessment that include scanning the log file or other auxiliary data structures, obtaining the precise damage profile can take significant amount of time. In order to reduce this time, we concentrated on making an estimated damage profile as soon as possible.

Our developed model is based exclusively on apriori knowledge of data relationships mined during normal database operation phase. The rules representing data relationships thus deduced help establishing the damage profile. Through experiment we concluded that with the increased data dependencies among data items in a database the accuracy of assessment increases steadily.

When an attacker or a malicious user updates a database, the resulting damage can spread to other parts of the database through valid users. A fast and accurate damage assessment must be performed as soon as such an attack is detected. We have developed two approaches for damage assessment in an affected database. While the first one uses transaction dependency relationships to determine affected transactions, the second approach considers data dependency relationships to identify affected data items for future recovery. These relationships are stored in a matrix format for faster manipulation. Our methods use pre-developed data structures to identify all affected transactions without requiring any log accesses. Since these data structures are built using bit-vectors and are manipulated using logical AND and OR operations, the damage assessment is done very quickly. Although the transaction dependency based model would work faster than the data dependency based model since the former approach accesses less number of bit-vectors requiring less processing, the latter can aid in precise damage assessment.

Using the above mentioned matrix, we also developed a technique for performing damage assessment in distributed systems. Furthermore, considering that transactions in a large database system may have little or no relationship with each other, a parallel damage assessment procedure is developed to further reduce the damage assessment time. In this approach, transactions are clustered based on the dependency relationships. Damage assessment is performed by finding the cluster that contains the malicious transaction and all other clusters that depend on the former. All clusters whose ancestor clusters are either unaffected or already checked can be processed in parallel. This dramatically reduced assessment time.

A database log is the primary resource for damage assessment and recovery after an electronic attack. The log is a sequential file stored in the secondary storage and it can grow to humongous proportions in course of time. To make the process of damage assessment and recovery more efficient, segmenting the log based on different criteria has been proposed before. But the trade off is that, either segmenting the log involves a lot of computation or damage assessment is a complicated process. In order to strike a balance between computation cost and complication with the damage assessment process, we developed hybrid log segmentation technique. Our method reduces the time taken to perform damage assessment while still segmenting the log fast enough so that no intricate computation is necessary. We built our model from a log that was previously segmented based on number of transactions, a time window for transactions to commit, or space occupied by committed transactions. While performing damage assessment, we re-segment the log based on transaction dependency. Thus during repeated damage assessment procedures, we create new segments with dependent transactions in them so that the process of subsequent damage assessment becomes faster when there are repeated attacks on the system. We have analyzed various cases that are applicable to the system and also developed algorithms for each of the cases. The algorithms are tested using a simulation model and the results showed significant reduction in damage assessment time compared to previously developed methods.

Indirect dependencies among transactions that are executed at various sites of a distributed database system make damage assessment and recovery in such a system a complicated process since it requires collaborations among multiple participant sites. Such collaborations are determined by the nature of global transactions, which have sub-transactions executed at multiple sites. Identification of affected transactions is usually the burden of each site manger, which is responsible for scanning the log and checking for the transaction dependency graphs. In order to make the process efficient, we have come up with two primary models, namely, centralized model and peer-to-peer model. The centralized model has been expanded to three different options. In the peer-to-peer model, each site communicates with a global coordinator extensively and the coordinator receives and forwards corresponding messages to the appropriate sites. The centralized model puts much of the burden of damage discovery on the coordinator site and, so, the requirement for message transmissions between each site and the coordinator is reduced.

Many critical information systems require that the database must remain available for transaction processing during and after detection of an attack. However, existing recovery schemes range from completely prohibiting access to affected databases to prohibiting access specifically to damaged data items. In case of critical information infrastructure this prohibition is undesirable. We have introduced a recovery model that can substantially reduce the database down-time during the recovery process, while attempting to provide repaired data as accurately as possible. The mechanism is divided into two activities. The first activity is the process of gathering summary information about all database attributes, which is an on going process throughout the lifetime of the database. The second activity occurs during and after an attack is recognized, providing acceptable values for damaged data items instantaneously, effectively keeping the database available at all times. Unlike prior recovery approaches our method does not block transactions that need access to damaged data items. Rather it provides them with acceptable fuzzy values of those data items when possible. This reduces denial-of-service dramatically.

Based on the concept described above, we came up with the notion of fuzzy dependency relationships in a database which can then be used in damage assessment and recovery process. We defined fuzzy dependency as a loose dependency relationship between two sets of attributes. It describes logical relationships among attributes in a database relation and those relationships can't be fully specified by functional dependencies, which focus on database schema and data organization. This characteristic of the database schema are then used to perform damage assessment and also to build fuzzy recovery model. Our developed model uses pre-derived rules to provide satisfactory values of damaged items. Querying fuzzy rules is faster than using other statistical tools and mathematical functions. This quick recovery is desired in many real-time applications, which do not require precise values of data items rather need faster data access.

With the goal of quickly identifying affected data items in case of an information attack without having to evaluate too many data items and then making the unaffected data items available as soon as possible we developed a mechanism based on information flow relationships. The model achieves faster damage assessment by analyzing the patterns of information flow within an organization (or closely related organizations), which we call a domain. A domain may have various types of service data. We build an information flow graph using data in many service groups. This graph is analyzed immediately after an attack detection to determine which services are affected. Then the remaining services and corresponding data items are made available to users while repair of damaged data continues. This model can be used along with other intrusion detection mechanisms. The effects of this model are largely dependent on the structure of a domain. If there are some “central points” in the probability (and/or information flow) graph and information flow tends to be in one direction, then the model has a better chance of eliminating large sets of data items, which are free of damages, at the earliest. In many cases, organizations have such structures with many central points, which acts as a critical nodes. Evaluating these central points first can point out the sets of data items that are damage free.

A new virus classification method, which groups computer viruses primarily based on their spreading mechanisms has been developed. By understanding the spreading methods of viruses it is hoped to effectively shut down or at least prevent viruses from infecting systems. This new virus classification technique is not intended to replace the current ones. Rather it investigates computer viruses from a different perspective and incorporates newly developed and potentially emerging viruses into a complete virus family. Resistances to each group of viruses can be developed based on their common characteristics. The developed method classifies not only existing viruses but also those highly potential viruses, which are technically possible at present and may appear in a reasonable future. The classification scheme has the following properties. Viruses with the same family genes are classified as one group. These family genes are mainly based on virus spreading techniques. Viruses fall into multiple sub families; but there is only one major group each virus belongs to. The classification is based on the way that a virus behaves or operates, not on the virus code.

In contrast to the traditional viruses, network-aware viruses have demonstrated new features and characteristics. In our effort, we have coined the term “network viruses”, which refers to those viruses spreading through networks. Security attacks can come from both viruses and hacking programs. But the combination of the two could be more powerful from the malicious perspective. Although these hacking empowered malicious programs borrow from hacking techniques, they are still computer viruses by nature since they pose the most fundamental features of computer viruses: replication, execution, and spreading. We surveyed several hundreds of compute viruses (most of them are network related) and classified them based on their spreading and infecting mechanisms. Virus intelligence is introduced to describe the various levels of implementing complexity and infecting abilities of network viruses. This concept also distinguishes the degrees to which different network viruses are related to hacking programs. A network virus makes use of networking protocols and/or applications to spread. It differs from a non-network virus in that it has built-in functions or library routines to take advantage of networking capabilities. Network viruses automatically break into other computers across the networks and then replicate their programs without user involvement. Like a traditional computer virus, a network virus typically consists of three modules (loading, spreading, and infecting modules) and one optional knowledge base.

Personnel Supported:

Faculty:

1. Thomas Wiggen (University of North Dakota)
2. Brajendra Panda (University of Arkansas)

Research Associate:

3. Arnada Pany

Graduate Students:

4. Yi Hu (Completed M.S. in Computer Science and is currently a Ph.D. student of B. Panda)
5. Yanjun Zuo (Completed M.S. in Computer Science and is currently a Ph.D. student of B. Panda)
6. Jing Zhou (Completed M.S. in Computer Science)
7. Prahalad Ragothaman (Started as a Ph.D. student, but did not complete)

Resulting M.S. Theses

1. "Data Dependency Approach to Database Intrusion Detection", Yi Hu, M.S., Computer Science, December 2002.
2. "Classification of Computer Viruses Based on Their Spreading Mechanisms", Yanjun Zuo, M.S., Computer Science, May 2003.
3. "An Implementation Oriented Approach for Database Damage Assessment", Jing Zhou, M.S., Computer Science, August 2003.

The above theses can be found at the University of Arkansas library.

Besides the above M.S. theses, Ph.D. dissertation work was started by the following students who were supported through the grant.

1. Yi Hu
2. Yanjun Zuo

Publications:

The following publications were resulted by Dr. Panda while he was partially supported by the project and as a direct or indirect outcome of this project.

Jing Zhou, Brajendra Panda, and Yi Hu, "Succinct and Fast Accessible Data Structures for Database Damage Assessment", In Lecture Notes in Computer Science, No. 3347, R. K. Ghosh and H. Mohanty (Editors) Springer Publications, Published as the Proceedings of the International Conference on Distributed Computing and Internet Technology (Systems Security Track), Bhubaneswar, India, December 2004.

Yanjun Zuo and Brajendra Panda, "Damage Assessment Models For Distributed Database Systems", In Research Directions in Data and Applications Security, XVIII, Csilla Farkas and Pierangela Samarati (Editors), p. 111-123, Kluwer Academic Publishers, Published as the Proceedings of the 18th Annual IFIP WG 11.3 Working Conference on Data and Application Security, Sitges, Spain, July 2004.

Yi Hu and Brajendra Panda, "Mining Data Relationships for Database Damage Assessment in a Post Information Warfare Scenario" In Proceedings of the 5th Annual IEEE Information Assurance Workshop, United States Military Academy, West Point, NY, June 9-11, 2004.

Yanjun Zuo and Brajendra Panda, "Fuzzy Dependency and Its Applications in Damage Assessment and Recovery", In Proceedings of the 5th Annual IEEE Information Assurance Workshop, United States Military Academy, West Point, NY, June 9-11, 2004.

Prahalad Ragothaman and Brajendra Panda, "Improving Damage Assessment Efficacy in case of Frequent Attacks on Databases," Data and Applications Security XVII: Status and Prospects, Sabrina De Capitani di Vimercati, Indrakshi Ray, and Indrajit Ray (Editors), p. 16-30, Kluwer Academic Press, 2004. (A detailed version of this paper appeared in Proceedings of 17th Annual IFIP WG 11.3 Working Conference on Data and Application Security, Estes Park, CO, August 4-6, 2003.)

Rajesh Yalamanchili and Brajendra Panda, "Transaction Fusion: A Model for Data Recovery from Information Attacks," Journal of Intelligent Information Systems, 23:3, pages 225-245, 2004.

Yi Hu and Brajendra Panda, "A Data Mining Approach for Database Intrusion Detection", In Proceedings of the 19th ACM Symposium on Applied Computing, Special Track on Database Theory, Technology, and Applications, Nicosia, Cyprus, March 14-17, 2004.

Yanjun Zuo and Brajendra Panda, "A Service Oriented System Based Information Flow Model For Damage Assessment", In Proceedings of the 6th IFIP WG 11.5 Working Conference on Integrity and Internal Control in Information Systems, Lausanne, Switzerland, November 13-14, 2003.

Brajendra Panda and Jing Zhou, "Database Damage Assessment Using A Matrix Based Approach: An Intrusion Response System," In Proceedings of the 7th International Database Engineering and Application Symposium, Hong Kong, July 16-18, 2003.

Yi Hu and Brajendra Panda, "Identification of Malicious Transactions in Database Applications," In Proceedings of the 7th International Database Engineering and Application Symposium, Hong Kong, July 16-18, 2003.

Amit Valsangkar and Brajendra Panda, "A Model for Making Data Available Ceaselessly During Recovery," In Proceedings of the 4th Annual IEEE Information Assurance Workshop, United States Military Academy, West Point, NY, June 18-20, 2003.

Yanjun Zuo and Brajendra Panda, "Network Viruses: Their Working Principles and Marriages with Hacking Programs," Poster paper, In Proceedings of the 4th Annual IEEE Information Assurance Workshop, United States Military Academy, West Point, NY, June 18-20, 2003.

Prahalad Ragothaman and Brajendra Panda, "Modeling and Analyzing Transaction Logging

Protocols for Effective Damage Assessment", Research Directions in Data and Applications Security, E. Gudes and S. Shenoi (Editors), p. 89-101, Kluwer Academic Press, 2003.

Satyadeep Patnaik and Brajendra Panda, "Transaction-Relationship Oriented Log Division for Data Recovery from Information Attacks", Journal of Database Management, Special issue on Data and Information Security, Volume 14, Number 2, April-June 2003.

Prahalad Ragothaman and Brajendra Panda, "Hybrid Log Segmentation for Assured Damage Assessment", In Proceedings of the 18th ACM Symposium on Applied Computing, Special Track on Database Systems, Melbourne, FL, March 10-12, 2003.

Yi Hu and Brajendra Panda, "Transaction Semantic Analysis for Database Intrusion Detection", 2002 Conference on Applied Research in Data Engineering, Little Rock, AR, Nov. 1, 2002.